基于多层感知机的蛋白质变性温度预测 *

丁雪松[a]，黄立群[a]，张步忠[a]，杨　洋[a, b†]，吕　强[a, b]

(苏州大学 a. 计算机科学与技术学院; b. 江苏省计算机信息处理技术重点实验室，江苏 苏州 215006)

摘　要：准确预测蛋白质变性温度在蛋白质工程和药物研制等领域具有重要意义。将全局特征和序列特征作为初始特征向量，利用提出的基于权值的降维算法对初始特征向量进行降维，降维后的特征输入多层感知机模型预测蛋白质变性温度。在盲测数据集上，该方法预测结果与实验测定结果的 PCC 值由降维前的 0.77 增加到 0.8，RMSE 值由降维前的 0.17 降低到了 0.16，蛋白质变性温度预测值的分类准确率与现有方法比较有明显提升。

关键词：蛋白质变性温度；多层感知机；回归预测

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2018.01.0134

Using multi-layer perceptron to predict protein melting temperature

Ding Xuesong[a], Huang Liqun[a], Zhang Buzhong[a], Yang Yang[a, b†], Lyu Qiang[A, b]

(*a. School of Computer Science & Technology Jiangsu Provincial Key Laboratory for Information Processing Technologies, Soochow University, Suzhou Jiangsu* 215006, *China*)

Abstract: It is significant to predict accurate protein melting temperature in protein engineering and drug design. In this paper, we proposed a novel weight-based dimensionality reduction algorithm, and applied it to obtain the input features of MLP model by using combination with global and sequential features as preliminary features. On blind test sets, the PCC value of predicted and experimental melting temperatures increased from 0.77 to 0.8, and RMSE value decreased from 0.17 to 0.16. The classification accuracy of predicted melting temperatures by our algorithm was significantly improved over the up-to-date service.

Key words: melting temperature of protein; multi-layer perceptron; regression prediction

0　引言

蛋白质稳定性是指蛋白质在高温环境下抵御热变性的一种能力，同时也是蛋白质保持自身最佳活性的固有特性。蛋白质变性温度是一个蛋白质功能是否丧失的一个重要衡量指标，是蛋白质动力学稳定性的度量方式之一，因此预测蛋白质变性温度在科学研究领域或药物研制等应用领域都有非常重要的意义。目前，蛋白质变性温度主要由差式扫描热量法(differential scanning calorimetry)、圆二色谱法(circular dichroism)、傅里叶变换红外光谱法(Fourier transform infrared spectroscopy)等实验方法测定，但是实验方法存在费用昂贵、流程复杂、周期长等不足。

近年来，利用数理统计[1]、机器学习[2]等方法预测蛋白质变性温度获得了广泛应用。Ku 等人[3]基于统计估算的方法，建立二肽含量与蛋白质变性温度之间的关联性来估算变性温度的范围，然而该方法不能预测蛋白质变性温度的具体数值。Pucci 等人[1]基于温度的统计势能预测同源蛋白质的稳定性曲线，该过程需要使用大量蛋白质的属性诸如蛋白质灵活度[4]、亲水性[5]、氢键[6]等需要实验测得，预测过程比较复杂。Gorania 等人[2]基于序列信息，构建人工神经网络和自适应模糊网络推理系统模型，通过分析蛋白质氨基酸序列和蛋白质变性温度之间的复杂非线性关系来预测蛋白质变性温度，然而该方法的数据量过小，并不能完全捕捉到蛋白质的特性与变性温度之间的关联。

近年来深度学习[7]在语音识别[8]、机器翻译[9]等领域的优异表现，受到人们的越来越多的关注与使用。本文基于多层感知机模型(multi-layer perceptron, MLP)对蛋白质变性温度进行预测，将全局特征和序列特征结合作为初始特征向量，利用基于权值的方法对特征进行降维。结果显示，本文方法在测试数据集上得到了均方根误差 0.16，皮尔森相关系数 0.8，对比实

收稿日期：2018-01-18；**修回日期：**2018-04-16　**基金项目：**国家自然科学基金项目（61170125，61602332）；江苏省高校自然基金项目（14KJB520035）资助，NSFC-广东联合基金（第二期）超级计算科学应用研究专项资助和国家超级计算广州中心支持

作者简介：丁雪松（1989-），男，河南信阳人，硕士研究生，主要研究方向为生物信息计算，深度学习（20154227019@stu.suda.edu.cn）；黄立群（1992-），男，江苏盐城人，硕士研究生，主要研究方向为生物信息计算，深度学习；张步忠（1980-），男，安徽安庆人，博士研究生，主要研究方向为生物信息计算，深度学习；杨洋（1981-），男（通信作者），江苏扬州人，博士，讲师，主要研究方向为生物信息学（yyang@suda.edu.cn）；吕强（1965-），男（通信作者），江苏苏州人，博导，教授，主要研究方向为生物信息计算，元启发搜索，并行计算（qiang@suda.edu.cn）.

验结果优于文献[3]。

1 材料与方法

1.1 数据集

本文的数据集来源于文献[10]，共有3520条蛋白质序列信息及其相应的全局变性温度，分别来源于四个组织：大肠杆菌E_coli 729条、酿酒酵母S. cerevisiae 709条，嗜热菌Thermus thermophilus 1073条和人类Human cervical cancer cells 1009条。首先按其分布比例抽取300条作为测试集，分别是E_coli:60条，S. cerevisiae:60条，Thermus thermophilus:90条，Human cervical cancer cells:90条，然后将其余3 220条蛋白质用作训练集。训练集与测试集具体条数如表1所示。

表1 训练集和测试集条数

数据集	训练集	测试集
大肠杆菌	669	60
酿酒酵母	649	60
嗜热菌	983	90
人类	919	90
总计	3220	300

1.2 特征提取与评估方法

1.2.1 全局特征

每条蛋白质抽取全局特征1 644维，分别是蛋白质结构和物化特性1 437维，基于氨基酸和蛋白质序列使用ProFEAT[11]计算得到；电子特性140维，基于原子的蛋白质电荷密度使用Protein_recon[12]得到；蛋白质功能和三维特征19维基于氨基酸序列使用ProtDCal[13]得到；蛋白质长度，相对分子质量以及各个氨基酸的数量与所占比例等48维，使用ExPASy[14]得到。

1.2.2 序列特征

本文所使用的序列特征包含两个部分：氨基酸分类和二肽键信息。根据物化特性将上述数据集中每条蛋白质的所有氨基酸分为6类[15]：疏水性(V, I, L, F, M, W, Y, C)、带负电荷(D, E)、带正电荷(R, K, H)、构象特殊(G, P)、极性(N, Q, S)、其他(A, T)。将每条蛋白质中上述6类氨基酸的数量和所占比例作为12维特征。本文将20种氨基酸以外的未知氨基酸称为X。最后基于20种氨基酸和X对蛋白质的二肽键数量和所占比例进行统计，构建882维特征。

1.2.3 初始特征向量

本文方法中采用的特征将上述全局特征和序列特征进行拼接，得到共2538维初始特征向量。本文对所有特征和标签全部采取了归一化处理。

1.2.4 基于权值的降维算法

机器学习中维度过高会导致较高的时间复杂度和空间复杂度，有时候过高的特征反而会带来噪音或者造成特征冗余，从而降低了准确率。另外，生物中某些特征的提取需要耗费大量的人力，财力，因此有必要进行一定程度的降维，过滤噪声，有更好的泛化性能，进而有利于提高精度和减少特征数，节省大量时间和金钱。

机器学习领域中数据降维是指采用某种映射方法，将原高维空间中的数据点映射到低维度的空间中。Principal component analysis(PCA)是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性。虽然PCA降维能够很好的降低维度，但是它的适用场景局限在线性降维中，而本文采用的是多层感知机预测蛋白质变性温度，是一种非线性回归模型，不同于单纯的数据降维，它们是一种特征选择，而非特征提取，是从数据源头开始分析以减少特征，从而为生物实验提取特征减少时间和金钱，且本文中提出的基于权值的降维方法对于多层感知机有很好的适用性。

本文通过构建一个MLP模型进行拟合，然后输出每个输入节点与后面隐层之间的权值关系，我们认为权值不论是正，是负，只要绝对值比较大，就认为对该神经网络起到了积极的作用（负值绝对值大是强抑制），因此本文对每个节点相关的权值集合中每个数绝对值化后进行阈值(0.0285)判断，计数超过阈值的总数过半即大于10（第一层隐层节点数为20），我们就认为这个节点输入的特征对变性温度有重要作用予以保留，否则进行剔除。最后本文将算法1应用于上述初始特征向量，得到了一个541维特征向量。然后重新建立一个MLP回归模型，输入基于上述方法筛选之后的特征，重新进行回归训练。算法1给出了基于权值降维的算法。

算法1 基于权值进行特征降维

```
Step1:  Build MLP model
Step2:  Input 2538 features to fit protein melting temperature
Step3:  Outputs the weight matrix
        For i=1 to 3220(numbers of protein):
            count=0
            For j=1 to 2538(numbers of protein features):
                If abs(weight matrix)>0.0285:
                    count+=1
            If count>10:
                Save the feature index
Step4:  From step 3 get features index matrix. According it, select new features for every protein.
        Build new MLP model.
        Input the new protein features to new MLP model to fit the protein melting temperature
```

1.2.5 评估方法

均方根误差(root mean square error,RMSE)，它是观测值与真实值偏差的平方和观测次数 n 比值的平方根。本 $X_{predicted,\,i}$ 是第 i 条蛋白质的模型预测值。n 是蛋白质的条数。

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(X_{observed,\,i}-X_{predicted,\,i})^2}{n}}$$

皮尔森相关系数(Pearson correlation coefficient,PCC)，是一种线性相关系数。皮尔森相关系数是用来反映两个变量线性相关程度的统计量。其中 N 为样本量，X 与 Y 分别表示样本的观测值与真实值。PCC 描述的是两组变量间线性相关强弱的程度。其绝对值越大表明相关性越强。本文中 X 与 Y 分别表示蛋白质的变性温度的观测值与预测值。N 表示的是蛋白质的条数。

$$PCC=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})(\sum Y^2-\frac{(\sum Y)^2}{N})}}$$

1.3 模型及训练

本文是基于 sklearn 平台搭建对蛋白质变性温度进行回归分析的 MLP[15~17]模型，如图 1 所示。本文使用的模型共构建了 3 层隐层，隐层节点数分别设置为(20,20,20)，激活函数设置为 relu。

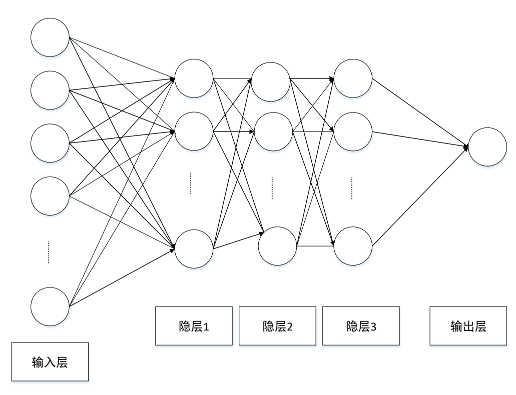


图 1 多层感知机模型图

多层感知机采用多隐层处理，比较适合进行非线性拟合函数，采用 BP 反向传播算法[18]，通过调节学习率，避免陷入局部最优解，该模型使用梯度下降算法降低 loss(损失函数)为优化目标。其中损失函数为

$$C(w,b)=\frac{1}{2n}\sum_{x}\left\|y(X)-a\right\|^2$$

优化目标为确定 w(权值)和 b(偏置)使得损失函数 C(w,b)最小，这意味着网络输出的值会越来越接近真实值。其中权值和偏置迭代公式如下：

$$\begin{cases} w_k = w_k-\dfrac{\eta}{m}\sum_j\dfrac{\partial C_{Xj}}{\partial w_k} \\ b_j = b_j-\dfrac{\eta}{m}\sum_j\dfrac{\partial C_{Xj}}{\partial b_j} \end{cases}$$

通过不断的迭代搜索到最合适的权值和偏置使得 loss 最小。

本文中利用蛋白质的特征作为输入数据输入到多层感知机的输入层，经过隐层进行非线性拟合，最后在输出层输出蛋白质的变性温度。由于本文提取蛋白质特征数据较多，特征之间可能存在关联，从而不利于模型进行训练以及影响模型的预测准确率。因此，本文使用算法 1，对特征向量进行降维。

2 结果

2.1 结果分析

本文首先使用 1.2.3 中初始特征向量输入到 1.3 中所述 MLP 模型进行蛋白质变性温度预测，得到测试集结果 PCC:0.772347，RMSE:0.1874。

对特征向量进行降维后得到 541 维特征，重新建立 MLP 模型进行回归拟合，得到的测试集结果 PCC:0.80559, RMSE:0.1638。特征向量降维前后效果如表 2 所示。

表 2 特征向量降维前后效果对比

特征向量维度	PCC	RMSE
2538	0.772347	0.1874
541	0.80559	0.1638

本文测试集中部分蛋白质变性温度预测值与对应的真实值对比如表 3 所示。

表 3 部分实验结果

蛋白质名称	预测值	真实值
Q5SJT3	80.1207	80.1028
P00950	51.0955	51.0743
Q5SJ36	82.1055	82.0235
P0CX47	54.0327	53.9060
Q5SLQ1	84.5149	84.2663
Q3E754	50.6029	50.2140
O43660	59.9184	59.4236
Q07551	53.6496	53.1438
Q5SH50	84.1124	83.5612
Q9UMS4	54.4323	49.3990
P30750	51.6206	46.4856
Q5SKM3	84.0083	75.4672
P0AGJ5	53.7757	43.6101

图 2 左图是降维前 2538 维特征的拟合图，图 2 右图是降维后 541 维特征的拟合图，可以看出前者较分散，后者较好的聚集于 y=x 上。

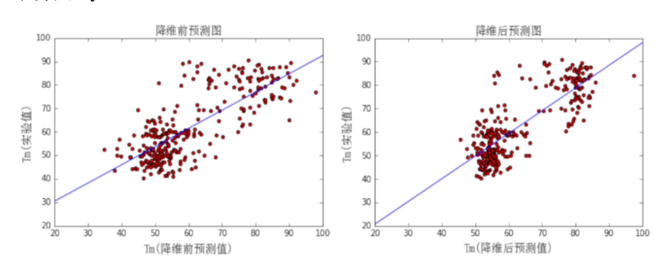


图 2 降维前后拟合效果对比图

横坐标是蛋白质变性温度预测值，纵坐标是蛋白质温度实验值。中间直线函数 y=x

2.2 对比实验

现有计算方法预测 Melting Temperature 的文献中，仅有 Ku[3] 提供了 webservice(http://tm.life.nthu.edu.tw/),本文提交了 1.1 所述测试集到该网站上进行预测，由于该方法只提供了对蛋白质变性温度的分类预测，分为(>65℃),(55℃~65℃),(<55℃)三类，因此本文基于分类准确率与之比较。其中由蛋白质的变性温度的实验值所属分类作为基准类别，对分类结果进行统计并评估。本文 1.2.4 中方法与 Ku 方法预测蛋白质变性温度的分类准确率对比如表 4 所示。

表 4 分类准确率对比

	准确数目	总数	准确率
Ku	114	300	0.38
MLP	189	300	0.63

3 结束语

本文从已知蛋白质的变性温度为目标进行拟合一个预测模型，以此来预测更多未知变性温度的蛋白质，意义在于为生物工程提供辅助依据，从而降低生物实验的时间和经济成本。本文基于 MLP,采用 2 538 维特征向量作为初始特征向量，利用权值筛选方法进行降维后，得到 541 维特征，取得了更高的预测性能。对比实验结果表明，本文所提出的预测模型不仅可以预测具体的蛋白质变性温度数值，应用在分类预测上，也比已报道方法表现更出色。尝试找到更具代表性的特征属性是提高预测蛋白质变性温度的难点，因此如何挖掘到这些属性是下一步工作的重点。

参考文献：

[1] Pucci F, Rooman M. Stability curve prediction of homologous proteins using temperature-dependent statistical potentials [J]. PLoS Computational Biology, 2014, 10 (7): e1003689

[2] Gorania M, Seker H, Haris P I, *et al.* Predicting a protein's melting temperature from its amino acid sequence [C]// Proc of Annual International Conference of Engineering in Medicine and Biology Society. 2010: 1820-1823.

[3] Ku Tienhsiung, Lu Peiyu, Chan Chenhsiung, *et al.* Predicting melting temperature directly from protein sequences [J]. Computational Biology & Chemistry, 2009, 33 (6): 445-450.

[4] Vihinen M. Relationship of protein flexibility to thermostability [J]. Protein Engineering, Design & Selection, 1987, 1 (6): 477-480.

[5] Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions [J]. Proteins, 1994, 19 (2): 141-149.

[6] Prevost M, Wodak S J, Tidor B, *et al.* Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96 (Ala mutation in barnase [J]. Proceedings of the National Academy of Sciences of the USA, 1991, 88 (23): 10880-4.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504-507.

[8] Graves A, Mohamed A, Hinton G E. Speech recognition with deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.

[9] Sutskever I, Vinyals O, Le Quoc V. Sequence to sequence learning with neural networks [J/OL]. (2014-10-14) . https://arxiv. org/pdf/1409. 3215. pdf.

[10] Leuenberger P, Ganscha S, Kahraman A, *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability [J]. Science, 2017, 355 (6327): eaai7825.

[11] Zhang Peng, Tao L, Zeng Xian, *et al.* PROFEAT update: a protein features web server with added facility to compute network descriptors for studying Omics-derived networks [J]. Journal of Molecular Biology, 2016, 429 (3): 416.

[12] Zaretzki J, Bergeron C, Rydberg P, *et al.* RS-Predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4 [J]. Journal of Chemical Information & Modeling, 2011, 51 (7): 1667-1689.

[13] Ruiz-Blanco Y B, Paz W, Green J, *et al.* ProtDCal: a program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins [J]. BMC Bioinformatics, 2015, 16 (1): 162.

[14] Gasteiger E, Hoogland C, Gattiker A, *et al.* Protein identification and analysis tools on the ExPASy server [M]// The Proteomics Protocols Handbook. [S. l.] : Humana Press, 2005: 531.

[15] Yang Yang, Abhishek N, Shen Bairong, *et al.* PON-Sol: prediction of effects of amino acid substitutions on protein solubility [J]. Bioinformatics, 2016, 32 (13): 2032-2034

[16] 王娟，吴宪祥，曹艳玲. 基于差分进化生物地理学优化的多层感知器训练方法 [J]. 计算机应用研究，2017, 34 (3): 693-696. (Wang Juan, Wu Xianxiang, Cao Yanling. Multi-layer perceptron using hybrid differential evolution and biogeography-based optimization [J]. Application Research of Computers, 2017, 34 (3): 693-696.)

[17] Kruse R, Borgelt C, Klawonn F, *et al.* Multi-layer perceptrons [M]// Computational Intelligence. London: Springer, 2013: 47-81.

[18] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [J]. Journal of Machine Learning Research, 2010, 9: 249-256.